Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 41 (2008) 2237–2252
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

# Human action recognition using shape and CLG-motion flow from multi-view image sequences☆

Mohiuddin Ahmad, Seong-Whan Lee*

*Department of Computer Science and Engineering, Korea University, Anam-dong, Seongbuk-ku, Seoul 136-713, Republic of Korea*

Received 26 June 2007; received in revised form 6 November 2007; accepted 4 December 2007

**Abstract**

In this paper, we present a method for human action recognition from multi-view image sequences that uses the combined motion and shape flow information with variability consideration. A combined local–global (CLG) optic flow is used to extract motion flow feature and invariant moments with flow deviations are used to extract the global shape flow feature from the image sequences. In our approach, human action is represented as a set of multidimensional CLG optic flow and shape flow feature vectors in the spatial–temporal action boundary. Actions are modeled by using a set of multidimensional HMMs for multiple views using the combined features, which enforce robust view-invariant operation. We recognize different human actions in daily life successfully in the indoor and outdoor environment using the maximum likelihood estimation approach. The results suggest robustness of the proposed method with respect to multiple views action recognition, scale and phase variations, and invariant analysis of silhouettes.


*Keywords:* Action recognition; Action matrix; Combined local–global (CLG) optic flow; Invariant Zernike moments; Multi-view image sequence; Multidimensional hidden Markov model (MDHMM)

## 1. Introduction

Recognition of human actions from multiple views image sequences is very popular in the computer vision community since it has applications in video surveillance and monitoring, human–computer interactions, model-based compressions, augmented reality, and so on. The existing methods of human action recognition can be categorized depending on the image state properties, such as motion-based, shape-based, gradient-based, etc. Several human action recognition methods have been proposed in the last few decades. Detailed surveys can be found in Refs. [1–4], where different methodologies of human action recognition, human movement, etc., are discussed. Based on these reviews, researchers either use human body shape information or motion information with or without body shape model for action recognition. Our approach can be considered as a combination of shape- and motion-based representation without using any prior body shape model.

One standard approach for human action recognition is to extract a set of features from each image sequence frame, and use these features to train classifiers and to perform recognition. Therefore, it is important to answer the following question. Which feature is robust to action recognition in critical conditions or varying environment? Usually, there is no rigid syntax and well-defined structure for human action recognition available. Moreover, there are several sources of variability [30] that can affect human action recognition, such as variation in speed, viewpoint, size and shape of performer, phase change of action, and so on, and the motion of the human body is non-rigid in nature. These characteristics make human action recognition a more challenging and sophisticated task. Considering the above circumstances, we consider some issues that affect the development of models of actions and classifications, which are as follows:

- The trajectory of an action from different viewing directions is different; some of the body parts (part of hand, lower part

☆ A preliminary version of the paper has been presented in the 7th IEEE International Conference on Automatic Face and Gesture Recognition, Southampton, UK, April 2006.

* Corresponding author. Tel.: +82 2 3290 3197; fax: +82 2 926 2168.
*E-mail addresses:* mohi@image.korea.ac.kr (M. Ahmad), swlee@image.korea.ac.kr (S.-W. Lee).

0031-3203/$30.00 © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2007.12.008

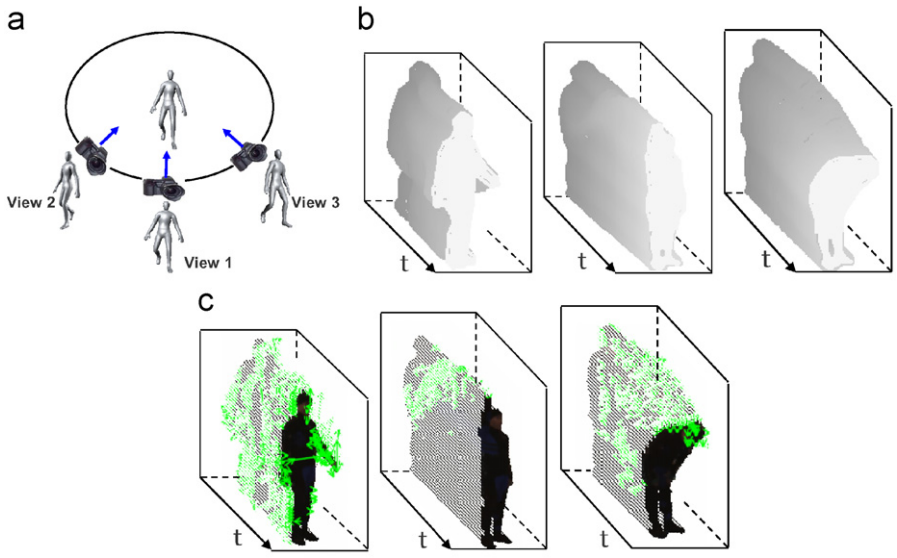


Fig. 1. Representation of human action using shape and motion sequences with multiple views. (a) Multiple views variation of an action. (b) Shape sequences (walking, raising the right hand, and bowing). (c) Motion sequences (walking, raising the right hand, and bowing). The motion distribution is different for each action.

of leg, part of body, etc.) are occluded due to view changes, which are shown in Fig. 1.

- An action can be viewed as a series of silhouette images of the human body (Fig. 1(b)). The silhouette information involves no translation, rotation, and scaling. Moreover, the silhouette sequence of an action is invariant to the speed.
- Action can be viewed by the motion or velocity of human body parts (Fig. 1(c)). Simple action involves the motion of a small number of body parts and complex action involves the motion of a whole body. The motion is non-rigid in nature.
- Human action depends on anthropometry, method of performing the action, phase variation (starting and ending time of the action), scale variation of an action, and so on.

Among various features, the motion or velocity of the body parts and human body shape play the most significant roles for recognition. Motion-based features can portray the approximation of the moving direction of the human body, and human action can be effectively characterized by motion rather than other cues, such as color, depth, and spatial features. In the motion-based approach, the motion information of the human such as optic flows, affine variation, filters, gradients, spatial–temporal words, and motion blobs are used for recognizing actions. Motion-based action recognition has been performed by several researchers, such as [6–17]. However, most motion-based techniques are not robust in capturing velocity when motions of the actions are similar for the same body parts. As an example, motion-based features can easily discriminate between walking and sitting down, but fail to discriminate between walking and slow running or jogging. On the other hand, the human body silhouette represents the pose of the human body at any instant in time, and a series of body silhouette images can be used to recognize human actions correctly, regardless of the speed of movement. Different descriptors of shape information of motion regions such as points, boxes, silhouettes, and blobs are used for recognizing or classifying actions. Several researchers performed action recognition using shapes or silhouettes, such as [16,18–24]. Therefore, combing shape and motion-based features overcome the limitations of either motion or shape-based behaviors. During the action recognition of persons, we utilize the combined motion and shape information for recognizing the periodic as well as non-periodic or single occurrence actions. Moreover, most of the human action recognition techniques depend on the viewing direction. However, the trajectory of an action from different viewing angles is different. The work of testing an action using multi-view motion learning is not well resolved. Seitz and Dyer in Ref. [15] described an approach to detect cyclic motions that is affine invariant. Rao and Shah [25] again used view invariant actions by affine invariance assuming that 2D positions of the hand are already known. This approach utilized spatiotemporal curvature maxima as instants of interest to map an unknown viewpoint to a "normal" viewpoint. The action was considered as being completely represented by the motion of the hand alone. In Ref. [26], authors presented human action in video using 3D model-based invariants and represent each action using a unique curve.

Fig. 2 shows a block diagram of the proposed method. In the preprocessing steps, the foreground is extracted by using background modeling, shadow elimination and morphological operation. From the foreground image, the velocity of an action is estimated by using combined local–global (CLG) optical flow. The global shape flow features are extracted from silhouette image sequence. The shape flow represents the flow deviation and invariant moments. We use the modified Zernike moment, which is robust against noise and invariant to scale, rotation, and translation, is used to reduce noise and to normalize the action data spatially. Motion features are extracted based on the same

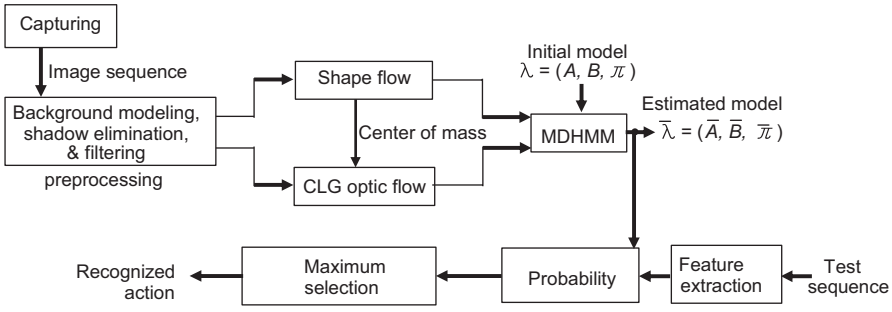


Fig. 2. Flow diagram of the proposed method.

center of mass (CM) of corresponding silhouette image. The combined features are then feed to multidimensional hidden Markov model (MDHMM). In the classification stage, matching of an unknown sequence with a model is done through the calculation of the probability that the MDHMM could generate the particular unknown sequence. The MDHMM with the highest probability most likely generated that sequence. The actions modeling and classification in this work involves both the Korea University Gesture database (KUGDB) [27] and the KTH database (KTHDB) [6]. In our work, for modeling and classifying human actions, the following techniques are proposed for extracting features from image sequences and classify actions:

- 2D Cartesian representation of CLG optic flow velocity vectors (horizontal and vertical component) are derived from the normalized flow of each quadrant (the origin of quadrant axis is the CM position of corresponding silhouette image), characterizing the recognition of motion in action with less noise.
- The shape flow features are extracted from the global flow of the shape. The global flow of human body silhouette images is extracted by applying robust description of geometric-orthogonal moments, flow deviations and anthropometry flow, with recognition of shapes in action.
- Learning of the combined features using the MDHMM in different viewing angles characterizes the recognition as view invariant.
- Normalization of any body shape to a suitable representation characterizes the anthropometry variation of performers.
- Translation, rotation, and scale invariant analysis of Zernike moments characterize view-invariant behavior of spatial action data and is used to reduce noise.
- The sources of variability of human actions, such as view-variation, phase variation of actions, and person's anthropometry are adapted to the system.

Therefore, based on the combined information of silhouettes, optical flows, sources of variabilities (such as view-variation, speed of actions, phase variation of actions, and person's sizes), and multiple views, human action recognition is more robust. We propose to recognize several actions of humans from multiple views learning of shape and motion features using the MDHMMs, since we use multiple features at the same time.

This paper is organized as follows: Section 2 presents action representation in our system. Section 3 briefly summarizes the basic algorithm of foreground extraction from the background. Section 4 discusses feature extraction using the shape flow and CLG motion flow. Section 5 describes and illustrates MDHMMs for action recognition. Section 6 presents experimental results and discussions of the selected approaches. Finally, conclusions are drawn in Section 7.

## 2. Human actions representation

### 2.1. Action in world and image coordinate system

Human action is the movement of humans for performing a task within a short period of time. The action may be simple or complex depending on the number of body limbs involved in the action. We consider that a complete human action representation might be the set of all 3D points on a performing actor. Therefore, we can consider human actions as 4D points in real world-space, which can be represented as follows:

$$\mathbf{A}_{4D} = \begin{pmatrix} X_j^{T_1} & X_j^{T_2} & \cdots & X_j^{T_p} \\ Y_j^{T_1} & Y_j^{T_2} & \cdots & Y_j^{T_p} \\ Z_j^{T_1} & Z_j^{T_2} & \cdots & Z_j^{T_p} \end{pmatrix}, \tag{1}$$

where $X$, $Y$, and $Z$ represent the state-space representation of a point in the 4D plane of a person performing an action. Here, $j$ represents the points set, or anatomical landmarks points, or voxel and $T_i$ is the $i$th frame in the world coordinate. When the human action is projected into the spatial–temporal image plane, then Eq. (1) can be represented by

$$\mathbf{A}_{3D} = \begin{pmatrix} x_j^{t_1} & x_j^{t_2} & \cdots & x_j^{t_p} \\ y_j^{t_1} & y_j^{t_2} & \cdots & y_j^{t_p} \end{pmatrix}, \tag{2}$$

where $x$ and $y$ represent the 2D points in the spatiotemporal image space or 3D space and $j$ represents the points set of pixel in the action region and $t_i$ is the $i$th frame in the image coordinate in performing an action. The relation between $t$ and $T$ could be linear, such that $t = \alpha_t T + d_t$, where $\alpha_t$ and $d_t$ represent the temporal coefficient and temporal constant, respectively.

### 2.2. Actions representation in our system

In our approach, human action from an image sequence, $f(x, y, t)$ is represented by a set of multidimensional CLG optic flow feature vector and shape flow feature vector. Therefore, the action matrix within the action boundary represents

$$\mathbf{A}_{system} = \begin{pmatrix} s_i^{t_s} & s_i^{t_s+1} & \cdots & s_i^{t_e} \\ v_j^{t_s} & v_j^{t_s+1} & \cdots & v_j^{t_e} \end{pmatrix}, \tag{3}$$

where $v_j$ represents the CLG optic flow velocity and $s_i$ represents the shape flow feature of an action with in the boundary of starting frame $t_s$ and ending frame $t_e$ with the period or duration $p = t_e - t_s$. Here, $s_i^{t_s} = \{s_1^{t_s}, s_2^{t_s}, \ldots, s_{D_1}^{t_s}\}$ and $v_j^{t_s} = \{v_1^{t_s}, v_2^{t_s}, \ldots, v_{D_2}^{t_s}\}$. Therefore, the combined features due to motion and shape at each frame is $D_1 + D_2$.

### 2.3. Diversity of action representation

To consider the diversity of modeling and classifying actions, we consider the anthropometry and phase variation of the action. For anthropometry variation, the image sequences are changed into $f(x \pm a, y \pm b, t)$. In this situation, the representation is same as Eq. (3) but the features are different. The variable "phase change" refers to the action occurred at different starting and ending state. The starting and ending phase of an action depends on persons, time, style, and so on. For example, in the 'bowing' action, a person bends the waist at different angles from the reference position, i.e. from a standing position. Therefore, the spatial–temporal image sequence, $f(x, y, t)$ can be alternatively represented by $f(x, y, t - \phi)$ for phase variation, where $\phi$ represents the time delay or phase change of an action. The value of $t_e$ in Eq. (3) is replaced by $t_e - \phi$ for ending phase variation of an action. Similarly, $t_s$ is replaced by $t_s + \phi$ starting phase variation of an action. The camera view information such as zooming of the person, slanting motion, and rotation of human body can be modeled by using affine transformation, $g(x_a, y_a, t) = f(a_1x + a_2y + d_x, a_3x + a_4y + d_y, t)$, where, $a_i$ and $d_j$ are constants. Therefore, these factors authorize the diversity of modeling the actions.

## 3. Preprocessing

In the preprocessing steps, we extract foreground, eliminate shadow, and then apply filtering. We then define the action boundary from the foreground image sequence. Briefly, these are explained below:

### 3.1. Foreground extraction

#### 3.1.1. Background modeling

We use background subtraction to extract the foreground, since the background is relatively static for all image sequences. We adopt a simple background modeling technique such as multiple Gaussian background modeling, for foreground extraction. For each subsequent frame, $p_t = [p_R(t), p_G(t), p_B(t)]$, we assume independence among different color channels. Several background images are accumulated and we extract the mean, standard deviation, and variance of the background images. Let $\mu_R$, $\mu_G$, $\mu_B$ be the mean values, and $\sigma_R$, $\sigma_G$, and $\sigma_B$ be the standard deviation of the background images which are computed over $N$ frames, then, we extract the foreground according to

$$p(x_t) = \begin{cases} 1 & \text{if } \begin{cases} |p_R(t) - \mu_R| \geqslant 2\sigma_R & \text{or} \\ |p_G(t) - \mu_G| \geqslant 2\sigma_G & \text{or} \\ |p_B(t) - \mu_B| \geqslant 2\sigma_B \end{cases} \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

#### 3.1.2. Shadow elimination

After background subtraction, there still exits some noises in the foreground, such as motion shadow. Therefore, the shadow elimination method should be adopted. Horprasert et al. proposed in Ref. [28] a pixel-based segmentation model in RGB color space which decomposes each background value into its brightness $\alpha$ and chromaticity distortion $CD$. In this method, for a given pixel, the expected background value $E_t = [\mu_R, \mu_G, \mu_B]$ is computed from $N$ training frames representing the static background. For each subsequent frame $p_t$, brightness ($\alpha_t$) and chromaticity distortions ($CD_t$) from the background value are given by

$$\alpha_t = \frac{((p_R(t)\mu_R/\sigma_R^2) + (p_G(t)\mu_G/\sigma_G^2) + (p_B(t)\mu_B/\sigma_B^2))}{[\mu_R/\sigma_R]^2 + [\mu_G/\sigma_G]^2 + [\mu_B/\sigma_B]^2}, \tag{5}$$

$$CD_t = \sqrt{\left(\frac{p_R(t) - \alpha_t\mu_R}{\sigma_R}\right)^2 + \left(\frac{p_G(t) - \alpha_t\mu_G}{\sigma_G}\right)^2 + \left(\frac{p_B(t) - \alpha_t\mu_B}{\sigma_B}\right)^2}. \tag{6}$$

In the RGB space, the chromaticity distortion is the length of the perpendicular vector between a pixel value $p_t$ and the line joining the zero intensity point and the background value $\mu$. It is an indicator of how much the pixel color differs from the background color. During the training phase, the variation $b$ of the chromaticity distortion is evaluated,

$$b = \sqrt{\frac{\sum_{t=0}^{N-1} CD_t^2}{N}}, \tag{7}$$

and used to compute a normalized chromaticity distortion, $\hat{CD} = CD_t/b$. For a given scene, the threshold $\tau_{CD}$ is chosen according to the successful detection rate of the shadow. Pixels are then labeled background, foreground, cast shadow, or highlight. Background pixels have small normalized brightness distortion, and small normalized chromaticity distortion. A pixel is labeled as cast shadow or highlight if it has a small normalized chromaticity distortion and a lower (cast shadow) or higher (highlight) brightness value than the background value. Unclassified pixels are labeled as foreground. More specifically, a pixel is labeled as cast shadow if these two conditions are respected: $\hat{CD}_t < \tau_{CD}$ and $\alpha_{min} < \alpha_t < 1$.

#### 3.1.3. Filtering

After the above shadow elimination step, there may exist some small regions and noise. For further preprocessing, several morphological operations such as erosion, dilation, and

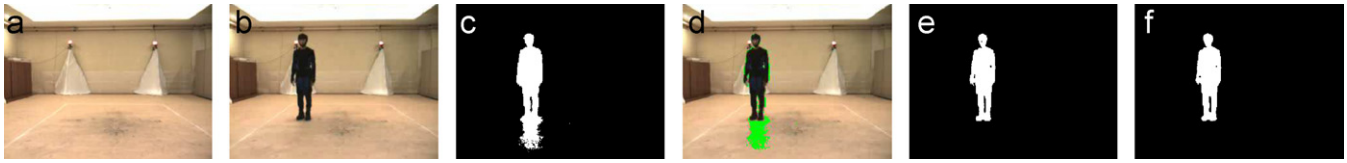


Fig. 3. Foreground extraction procedures. (a) Background image. (b) Current image. (c) Extracted foreground image with shadow. (d) Detected shadow pixels (green color). (e) Foreground image after shadow removal. (f) Foreground image after morphological and filter operations.

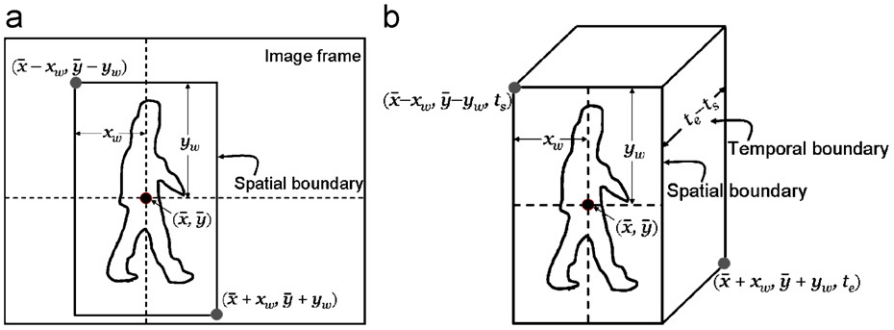


Fig. 4. Action boundary definitions. (a) Spatial boundary. (b) Action volume boundary by using spatial and temporal boundaries.

connected component analysis should be adopted. Finally, the resulting foreground image is obtained by median filtering. The neighboring window size of the median filtering is $5 \times 5$. Fig. 3 shows the image preprocessing steps for foreground extraction.

### 3.2. Action boundary

We define the action boundary as the action region in the image sequence where the movements of the person occur or the person exists. The action boundary depends on (1) anthropometry of human body, (2) distance between the video sensor and person performing action, and (3) type of action. Due to the above reasons, the action boundary of an image sequence varies and is extracted automatically from the filtered foreground image. The process is done according to the following steps:

- Find the CM of the silhouette $(\bar{x}, \bar{y})$ as shown in Fig. 4(a), where $(\bar{x}, \bar{y}) = (m_{10}/m_{00}, m_{01}/m_{00})$ and $m_{pq} = \sum_x \sum_y x^p y^q f(x, y)$.
- Define the maximum width and height distance from CM. Suppose they are $x_w$ and $y_w$.
- Now, define the spatial action boundary by a bounding box of corner points $(\bar{x} - x_w, \bar{y} - y_w)$ and $(\bar{x} + x_w, \bar{y} + y_w)$. It may be the same size of the original image.
- The temporal boundary of the action is bounded by starting time $(t_s)$ and ending time $(t_e)$.

According to the definition of action boundary, Fig. 4(a) shows the spatial boundary of the current frame and (b) shows the spatial–temporal boundary of the action. The spatial–temporal boundary is defined by two extreme points, $(\bar{x} - x_w, \bar{y} - y_w, t_s)$ and $(\bar{x} + x_w, \bar{y} + y_w, t_e)$. The spatial action boundary for all specified actions is assumed as the same and the temporal action boundary varies from action to action. For example, the temporal boundary among walking, jogging, and running can be described as $(t_e - t_s)_{running} \leqslant (t_e - t_s)_{jogging} \leqslant (t_e - t_s)_{walking}$. This is not fixed, because it depends on person and style.

## 4. Feature extraction

We use the CLG optic flow and shape flow feature for action representation and classification. The silhouette image sequence is used to extract the shape flow features and the foreground image sequence is used to extract the CLG motion flow features.

### 4.1. Shape flow

We define the shape flow as the global flow of silhouettes over the period of an action. The shape flow is characterized by the invariant geometric and Zernike moments, and flow deviations over the silhouette sequence, and global anthropometric variations. Therefore, the global motion of the shape can be integrated by multiple features of silhouette images, and can be stated by $s_i = [s_g, s_z, s_d, s_a]^{\mathrm{T}}$ where the symbols are defined and are described in the following subsections.

#### 4.1.1. Geometric moments

Moments and function of moments have been utilized as pattern feature in pattern recognition applications. Such features capture global information about the image and do not require close boundaries as required by Fourier descriptors. Hu [29] introduced seven nonlinear functions, $h_i$, where $i = 1, 2, \ldots, 7$ defined on regular moments using central moments which are translation, scale, and rotation invariant. These seven so called moment invariants were used in many pattern recognition problems. We use $s_g = [h_1, h_2, h_3, h_4]^{\mathrm{T}}$ as the geometric moment feature.

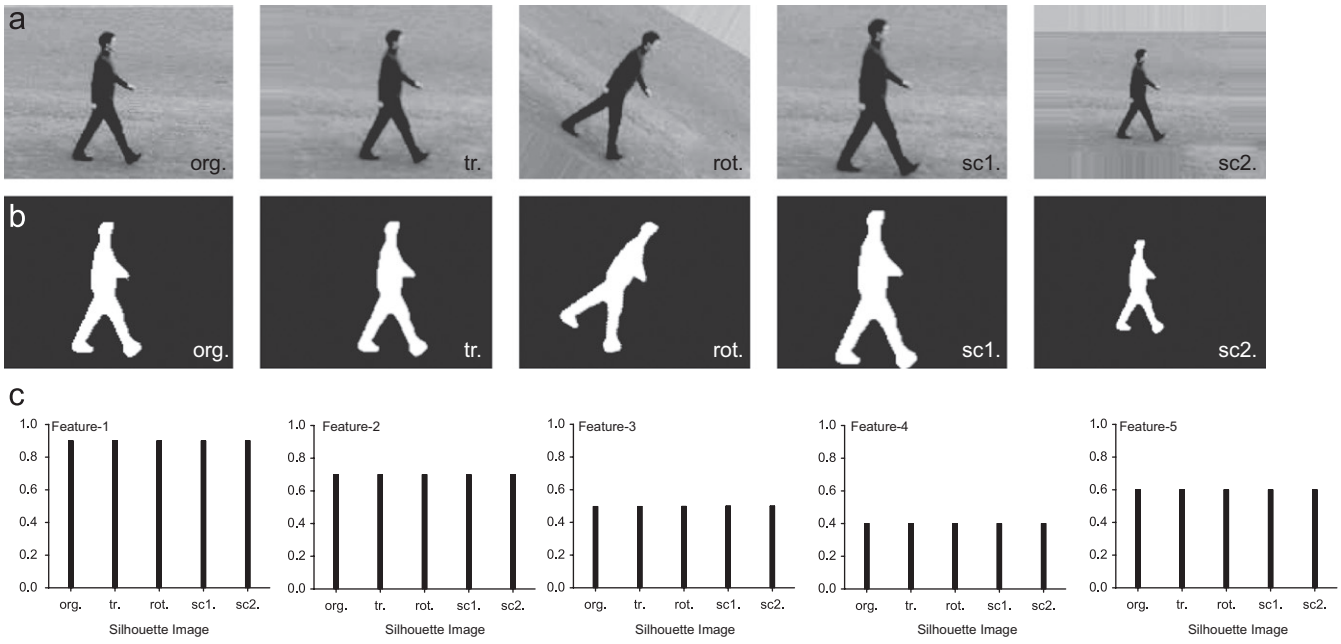


Fig. 5. Invariant analysis of geometric and Zernike moments of shape analysis. Each row represents the original (org.), translated (tr.), rotated (rot.), scaling-up (sc1.), and scaling-down (sc2.) images. (a) Original images; (b) silhouette images; (c) invariant features.

#### 4.1.2. Zernike moments

The geometric moment shows highly inaccurate results when the image is noisy. Zernike polynomials provide very useful moment kernels, present native rotational invariance and are far more robust to noise. Scale and translation invariance can be implemented using moment normalization. The magnitude of Zernike moments of the image sequence has been treated as global shape flow because Zernike moments are rotation invariant. We use the modified 2D Zernike moments of the silhouette images. The 2D Zernike moments of the image intensity function $f(\rho, \theta)$ with order $n$ and repetition $m$ is expressed as follows [30]:

$$Z_{nm} = \frac{n+1}{\lambda_N} \int_0^{2\pi} \int_0^1 V_{nm}^*(\rho, \theta) f(\rho, \theta) \rho \, \mathrm{d}\rho \, \mathrm{d}\theta, \tag{8}$$

where $V_{nm}^*(\rho, \theta) = R_{nm}(\rho) \exp(-\mathrm{j}m\theta)$ and $R_{nm}(\rho)$ is a radial polynomial defined by $R_{nm}(\rho) = \sum_{s=0}^{(n-|m|)/2} (-1)^s ((n-s)!/(s)!((n+|m|)/2-s)!((n-|m|)/2)!) \rho^{n-2s}$. Here $\rho = \sqrt{(2x-N+1)^2 + (N-1-2y)^2}/N$ with condition $0 \leqslant \rho \leqslant 1$, $\theta = \tan^{-1}(N-1-2y)/(2x-N+1))$, and $\lambda_N$ is a normalization factor. Here, $n$ is a non-negative integer and $m$ is a positive or negative integer subject to constraints $n - |m| = even$, and $|m| <= n$. For each silhouette image and one given value $\rho$ (length of vector from origin to $(x, y)$ pixel), we obtain the complex Zernike moments and is given by $Z_{nm} = ((n+1)/\lambda_N) \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} f(x, y) R_{nm}(\rho) \exp(-\mathrm{j}m\theta)$. The image position and scale is not fixed, therefore, the invariant analysis of the moment is necessary. To achieve scale and translation uniformity, the regular moments (i.e. $m_{pq}$) or radial polynomials of each image can be utilized. In general, an image function $f(x, y, t)$ can be normalized with respect to scale and translation by transforming it into $g(x, y, t)$ [30], where

$$g(x, y, t) = f\left(\frac{x}{a} + \bar{x}, \frac{y}{a} + \bar{y}, t\right), \tag{9}$$

with $(\bar{x}, \bar{y})$ being the CM of $f(x, y, t)$ and $a$ is the scale factor. The corresponding invariant Zernike moment $\hat{Z}_{nm}$ can be expressed by $\hat{Z}_{nm,t} = ((n+1)/\lambda_N) \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} g(x, y, t) R_{nm}(\rho) \exp(-\mathrm{j}m\theta)$. We exploit the concept and analysis of translation and scale invariants from [31,32] and use the translation and scale invariant features. We use $s_z = [\hat{Z}_{22}, \hat{Z}_{20}, \hat{Z}_{31}]^{\mathrm{T}}$ as the Zernike moment feature. We use the invariant Zernike moment features along with geometric features for robust description of shape flow. As an example, Fig. 5 shows the normalized invariant moment feature by using column plot for the shape description. Feature-1 ($h_1$) and feature-2 ($h_2$) show the invariant geometric moments and feature-3 ($\hat{Z}_{22}$), feature-4 ($\hat{Z}_{20}$), and feature-5 ($\hat{Z}_{31}$) show the invariant Zernike moment features. The percentage errors of invariants are less than 0.5%.

#### 4.1.3. Global motion deviation of silhouettes

We mentioned that the shape flow describes the global motion over the full time frame. For any silhouette image, the mean absolute deviation $\mathbf{d}_{kl}(t) = (d_{kx}(t), d_{ky}(t))$ from the CM in the direction of $x$ and $y$ of a silhouette $f(x, y, t)$ is used for shape flow description.

$$d_{kl,t} = \begin{cases} x-flow: \dfrac{\int_{(x,y)\in f(x,y,t) \geqslant Th} |x-\bar{x}| f(x,y,t) \, \mathrm{d}x \, \mathrm{d}y}{\int_{(x,y)\in f(x,y,t) \geqslant Th} f(x,y,t) \, \mathrm{d}x \, \mathrm{d}y}, \\ y-flow: \dfrac{\int_{(x,y)\in f(x,y,t) \geqslant Th} |y-\bar{y}| f(x,y,t) \, \mathrm{d}x \, \mathrm{d}y}{\int_{(x,y)\in f(x,y,t) \geqslant Th} f(x,y,t) \, \mathrm{d}x \, \mathrm{d}y}, \end{cases} \tag{10}$$

where $Th$ represents the threshold value and $Th = 0$. With this shape flow, we can distinguish between actions where more body parts are involved in motion (for example, sitting on floor,

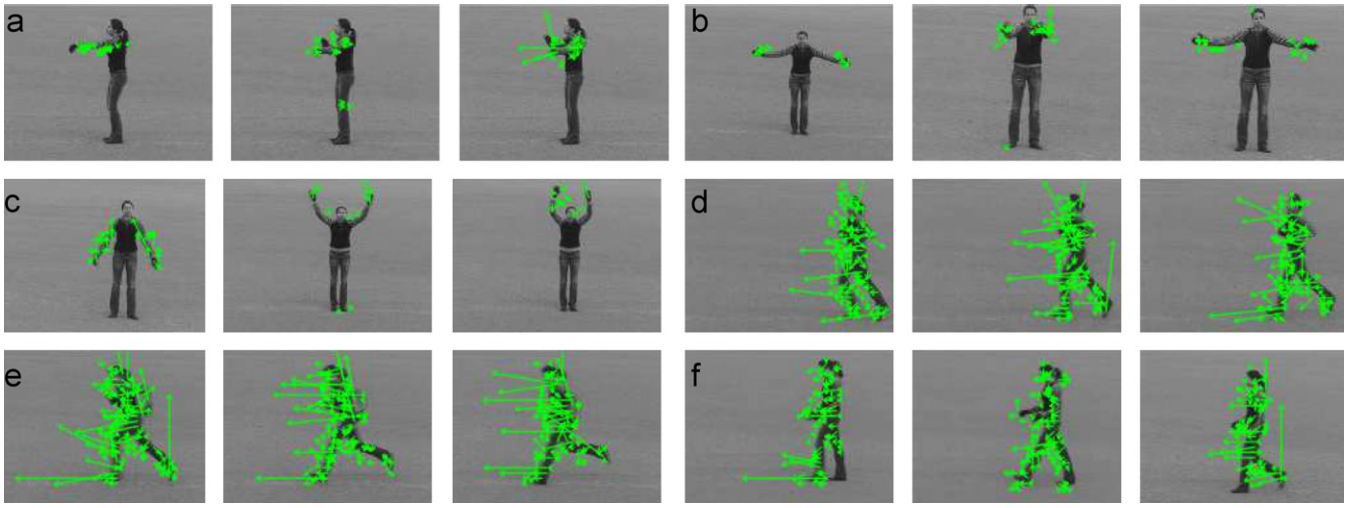


Fig. 6. CLG optical flows overlapping on the image of several actions in some selected frames (KTHDB). (a) Boxing; (b) hand clapping; (c) hand waving; (d) jogging; (e) running; (f) walking.

getting down on the floor, lying down on the floor, etc.), and an action concentrated in a smaller area where only small parts of the body move (for example, sitting on a chair, bowing, etc.). This flow feature can be considered as span or wideness of motion. Another important feature for describing global motion is the mean intensity of shape flow, $d_{ke}(t)$ of a silhouette $f(x, y, t)$, which represents the average absolute height or elevation of shape distribution, which can be expressed as

$$d_{ke,t} = \frac{\iint_{(x,y)\in f(x,y,t)\geqslant 0} f(x, y, t)\,\mathrm{d}x\,\mathrm{d}y}{\iint_{(x,y)\in f(x,y,t)\geqslant 0} \max f(x, y, t)\,\mathrm{d}x\,\mathrm{d}y}. \tag{11}$$

A large value of $d_{ke,t}$ indicates very intense flow of the silhouette and a small value indicates minimal flow. Therefore, the global flow deviations $s_d = [d_{kx}, d_{ky}, d_{ke}]^{\mathrm{T}}$ is used as the shape features. In addition to flow deviation, we use the global anthropometry variation of a person in the image sequence of an action as a feature. The projected width (proj.w) and height (proj.h) or their ratio of the person (i.e. silhouette) can be used to express the feature. We use $s_a = [f(x, y, t)_{proj.w}, f(x, y, t)_{proj.h}]^{\mathrm{T}}$ as the anthropometric shape flow.

### 4.2. CLG optic flow features

We use the CLG optic flow velocity as the motion feature, because it can precisely determine the motion. The CLG optic flow method proposed in Ref. [33] has complementary advantages over either global [34] or local [35] methods, due to its robustness against noise and combination of both local and global flow. For brief analysis of the method, the following notations are used: CLG optic flow: $\mathbf{v} = [v_x, v_y, 1]^{\mathrm{T}}$ at pixel $\mathbf{x} = (x, y)$, velocity gradient : $\nabla\mathbf{v} = |\nabla v_x|^2 + |\nabla v_y|^2$, intensity gradient : $\nabla_3 p = (p_x, p_y, p_t)^{\mathrm{T}}$, and motion tensor : $J_\rho(\nabla_3 p) = K_p * (\nabla_3 p \nabla_3 p^{\mathrm{T}})$. $K_p$ is smoothing kernel (spatial or spatial–temporal). The spatiotemporal version of the CLG functional is given by

$$E_{CLG}(\mathbf{v}) = \int_{video} (\mathbf{v}^{\mathrm{T}} J_p(\nabla_3\rho)\mathbf{v} + \alpha|\nabla_3\mathbf{v}|^2)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t, \tag{12}$$

where convolutions with Gaussians are now to be understood in a spatiotemporal way and $|\nabla_3\mathbf{v}|^2 = |\nabla_3 v_x|^2 + |\nabla_3 v_y|^2$. It minimizes the optical flow field, $\mathbf{v}(\mathbf{x}, t) = (v_x(\mathbf{x}, t), v_y(\mathbf{x}, t))$ using Euler–Lagrange equations [33]. The Euler–Lagrange equations are given by

$$\triangle_3 v_x - \frac{1}{\alpha}(J_{11}v_x + J_{12}v_y + J_{13}) = 0, \tag{13}$$

$$\triangle_3 v_y - \frac{1}{\alpha}(J_{12}v_x + J_{22}v_y + J_{23}) = 0, \tag{14}$$

where, $\alpha$ is the smoothing constant, $\triangle_3$ denotes the spatial–temporal Laplacian and $\triangle_3 = \partial_{xx} + \partial_{yy} + \partial_{tt}$. $J_{nm}$ is "motion tensor" or "structure tensor" which denotes the $(i, j)$th components of $J_\rho(\nabla_3 p)$ and is given by

$$J = \begin{pmatrix} J_{11} & J_{12} & J_{13} \\ J_{21} & J_{22} & J_{23} \\ J_{31} & J_{32} & J_{33} \end{pmatrix} = \begin{pmatrix} p_x^2 & p_x p_y & p_x p_t \\ p_y p_x & p_y^2 & p_y p_t \\ p_t p_x & p_t p_y & p_t^2 \end{pmatrix}. \tag{15}$$

The solution of Eq. (12) is given in Ref. [33] and thus we estimate the velocities $v_x(x, y, t)$ and $v_y(x, y, t)$. Fig. 6 shows the optical flow velocity overlapping on the image of several actions. It is found that related body parts involve optical flow velocity. For example, when the person conducts the "hand waving" action, motion only involves the hand. Similarly, when the person conducts the "running" action, then motion involves the whole body. For consistency of further analysis, the CLG flows are normalized at any instant of time, by

$$v_{n\mathbf{x}}(\mathbf{x}, t) = \begin{cases} x - flow\text{: } \frac{v_x(\mathbf{x},t) - v_{x.min}(t)}{(v_{x.max}(t) - v_{x.min}(t)) + \delta_m}, \\ y - flow\text{: } \frac{v_y(\mathbf{x},t) - v_{y.min}(t)}{(v_{y.max}(t) - v_{y.min}(t)) + \delta_m}, \end{cases} \tag{16}$$

where $v_{n\mathbf{x}}(\mathbf{x}, t)$ represents the normalized optical flow (the $x$-component or $y$-component velocity) in the spatial action

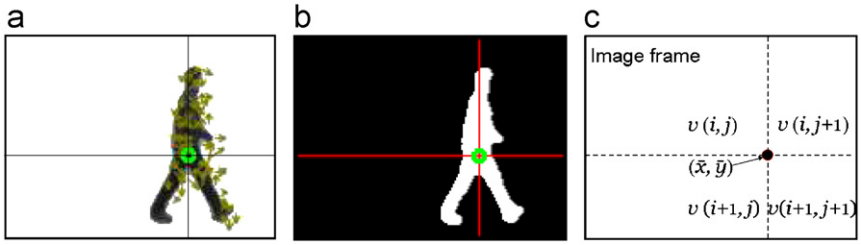


Fig. 7. Multi-geometry CLG motion flow features extraction. (a) CLG flows show in the quadrant regions. The small circle represents the CM of the image and the lines divided the image into quadrants. (b) The selection of CM position for (a). (c) Four quadrant blocks from the CM.

boundary. Moreover, $v_{x\cdot max}$ and $v_{x\cdot min}$ represent the maximum and minimum motion of $v_x(\mathbf{x}, t) \in I_a$, where $I_a$ is the spatial boundary of an action. Similarly, $v_{y\cdot max}$ and $v_{y\cdot min}$ represent the maximum and minimum motion of $v_y(\mathbf{x}, t) \in I_a$. We use the constant value $\delta_m$ for avoiding zero in the denominator. In order to extract the features from normalized flow, we partition the spatial action boundary into four quadrant blocks, $B(k)$ of equal size, as shown in Fig. 7. The four quadrants are described by (i) $\{(\bar{x} - x_w, \bar{y} - y_w), (\bar{x}, \bar{y})\}$, (ii) $\{(\bar{x}, \bar{y} - y_w), (\bar{x} + x_w, \bar{y})\}$, (iii) $\{(\bar{x} - x_w, \bar{y}), (\bar{x}, \bar{y} + y_w)\}$ and (iv) $\{(\bar{x}, \bar{y}), (\bar{x} + x_w, \bar{y} + y_w)\}$. The point $(\bar{x}, \bar{y})$ denotes the CM, $x_w$ is the width, and $y_w$ is the height of the block of the current silhouette image. Therefore, the flow feature vectors are extracted at each block with $n_B$ number of pixels using

$$v_{kl,t} = \begin{cases} x - flow: \frac{1}{n_B(pix>0)} \sum\limits_{\mathbf{x}\in B(k)} v_{nx}(\mathbf{x}, t), \\ y - flow: \frac{1}{n_B(pix>0)} \sum\limits_{\mathbf{x}\in B(k)} v_{ny}(\mathbf{x}, t), \\ abs.flow: \frac{1}{n_B(pix>0)} \sum\limits_{\mathbf{x}\in B(k)} \sqrt{v_{nx}^2(\mathbf{x}, t) + v_{ny}^2(\mathbf{x}, t)}. \end{cases} \quad (17)$$

Here, the vector $v_{kl,t}$ represents either the Cartesian component ($x$-component and $y$-component) of flow features or absolute CLG flow of the action boundary at any time. The subscript $l = \{x, y, abs.\}$, $k$ denotes the number of blocks, $pix$ represents the nonzero pixel value in the spatial boundary, and $n_B$ is the number of motion pixels at any block.

### 4.3. Combined shape and CLG flow

For each image frame in any action, the combined flow consists of shape flow and CLG motion flow. The combined flow can also be termed as the key features at any instant of time $t$ of the image sequence. The combined flow at time $t$ is given by $c_t = [s_i, v_j]^{\mathrm{T}}$. Each action video in any view direction or any scenario $d$ can be represented as an image sequence with starting time $t_s$ and ending time $t_e$. Therefore, features in the action volume boundary are expressed by

$$H_d = [c_{t_s,d}, c_{t_s+1,d}, \ldots, c_{t_e,d}], \quad (18)$$

where $(t_e - t_s)$ is the number of frames used in an action video. The value of $(t_e - t_s)$ depends on action variation and performer. For starting and ending phase variation of an action, Eq. (18) is modified according to

$$H_{d.phase} = \begin{cases} \text{start delay: } [c_{t_s+\phi(t_e-t_s),d}, c_{t_s+2\phi(t_e-t_s),d}, \ldots, c_{t_e-t_s}], \\ \text{end delay: } [c_{t_s,d}, c_{t_s+1,d}, \ldots, c_{t_e-t_s-\phi(t_e-t_s)}]. \end{cases} \quad (19)$$

## 5. Action modeling and classification using MDHMMs

The hidden Markov models have been widely used for analyzing time sequential data, such as speech recognition and online handwriting recognition. We organize the action recognition system using motion and shape flow feature and multidimensional hidden Markov models. Some previous research on gesture and human activity recognition used HMMs, such as in Ref. [36–40,16,41–43]. We choose data that consist of several independent components. Therefore, we use MDHMMs for modeling human actions. For multi-view recognition of human actions, we build an MDHMM model for each action.

### 5.1. Action modeling using MDHMMs

Before modeling human actions, we review the basic hidden Markov model (HMM) notation. Detailed explanation of the HMM may be found in several sources including [44].

- $S = \{S_1, S_2, \ldots, S_N\}$—a set of $N$ states. The state at time $t$ is denoted as $q_t$.
- $V = \{v_1, v_2, \ldots, v_M\}$— a set of $M$ distinct observation symbols. The observation at time $t$ is denoted as $o_t$.
- $A = a_{ij}|_{N\times N}$ is the state transition matrix whose elements $a_{ij} = P(q_{t+1} = S_j | q_t = S_i)$ are transition probabilities.
- $B = \{b_j(O_k)\}|_{N\times M}$ is the observation symbol probability matrix, where $\{b_j(O_k)\}$ is the probability of emitting $v_k$ at time $t$ in state $S_j$: $b_j(k) = P(\mathbf{o}_t = v_k | q_t = S_j)$.
- $\pi = \{\pi_i\}$—an initial state distribution where $\pi_i$ is the probability of the initial state: $\pi = P(q_1 = S_i) = [\pi_1, \pi_2, \ldots, \pi_N]$.

The complete parameter set of the HMM can be compactly expressed as

$$\lambda = \{A, B, \pi\}. \quad (20)$$

There are three key problems that must be solved for real application of the HMM: evaluation, decoding, and estimation.

To train the reference data of action sequences, the Baum–Welch algorithm is used. To obtain an HMM, we need to compute $P(O|\lambda)$. Since, we use the combined optic flow and shape flow vector for action recognition, then we can consider more observable symbols at each time $t$. Therefore, we choose the MDHMM which was proposed in Ref. [45] for skill learning to telerobotics. To deal with multidimensional data, the original HMM algorithms must be modified. For a $R$ dimensional HMM, in state $q_t = S_i$, $M \times R$ distinct output symbols

$$O = \{O_1, O_2, \ldots, O_M\}, \tag{21}$$

can be observed, where $R$ is the dimensions of the features in space and $O_k = [o_k(1), o_k(2), \ldots, o_k(R)]$. If we assume that each dimensional signal is stochastically independent, then based on this assumption, the forward variable $\alpha$ computation in Ref. [44] can be modified as follows:

$$\begin{cases} \textit{Initilization}: \alpha_1(i) = \pi_i \prod_{l=1}^{R} b_i(O_1(l)), \\ \textit{Induction}: \alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] \prod_{l=1}^{R} b_j(O_{t+1}(l)). \end{cases} \tag{22}$$

Similarly, the backward variable $\beta$ in Ref. [44] is computed as follows:

$$\left\{ \textit{Induction}: \beta_t(i) = \left[\sum_{j=1}^{N} a_{ij} \beta_{t+1}(j)\right] \prod_{l=1}^{R} b_j(O_{t+1}(l)) \right. . \tag{23}$$

An iterative algorithm is used to update the model parameters. Consider any model $\lambda$ with nonzero parameters. We first define the posterior probability of transitions, $\gamma_{ij}$ from state $i$ to state $j$, given the model and observation sequence,

$$\begin{aligned} \gamma_t(i, j) &= P(S_t = i, S_{t+1} = j | O, \lambda) \\ &= \frac{\alpha_t(i) a_{ij} \prod_{l=1}^{R} b_j(O_{t+1}(l)) \beta_{t+1}(j)}{P(O|\lambda)}. \end{aligned} \tag{24}$$

Similarly, the posterior probability of being in state $i$ at time $t$, $\gamma_t(i)$, given the observation sequence and model, is defined as

$$\gamma_t(i) = P(S_t = i | O, \lambda) = \frac{\alpha_t(i) \beta_t(i)}{\sum_{k=1}^{N} \alpha_T(k)}. \tag{25}$$

Here, $\sum_{t=1}^{T-1} \gamma_t(i)$ can be interpreted as the expected (over time) number of times that state $S_i$ is visited. Using the above formulas and the concept of count event occurrences, a new model $\bar{\lambda} = (\bar{A}, \bar{B}, \bar{\pi})$ can then be created to iteratively improve the old model $\lambda = (A, B, \pi)$. The Baum–Welch algorithm in Ref. [44] extends to the multidimensional case based on the previous independent assumption:

(1) state transition probability:

$$\bar{a}_{(i,j)} = \frac{\sum_{t=1}^{T-1} \gamma_t(i, j)}{\sum_{t=1}^{T-1} \sum_j \gamma_t(i, j)}, \quad 1 \leqslant i, \quad j \leqslant N, \tag{26}$$

(2) symbol emission probability:

$$\begin{aligned} \bar{b}_j^{(i)}(k) = \frac{\sum_{t \in O_t(i) = v_k^{(i)}} \gamma_t(j)}{\sum_t^{\mathrm{T}} \gamma_t(j)}, \quad 1 \leqslant i \leqslant R, \quad 1 \leqslant j \leqslant N, \\ 1 \leqslant k \leqslant M, \end{aligned} \tag{27}$$

(3) initial state probability:

$$\bar{\pi}_j = \gamma_1, \tag{28}$$

where $v_k^{(i)}$ are the observation symbols. If we repeat the above estimation and use $\bar{\lambda} = (\bar{A}, \bar{B}, \bar{\pi})$ to replace $\lambda$, it ensures that $P(O|\lambda)$ can be improved until a limiting point is reached. The Baum–Welch algorithm gives the maximum likelihood estimate of MDHMM and can be used to obtain the model that describes the most likelihood human action for given features.

### 5.2. Action classification using HMMs

We classified the image sequences manually into different classes and views. The trained model is used to classify the actions. The forward–backward algorithm or the Viterbi algorithm can be used to classify the actions from any specified view.

To recognize the input action sequences, the Viterbi algorithm [44] is used. This decoding problem is how to find the best state sequence, given that an observation sequence $O = \{O_1, O_2, \ldots, O_T\}$ and a model $\lambda = (A, B, \pi)$.

$$q = (q_1, q_2, \ldots, q_T), \tag{29}$$

where $q_t$ is the actual state at time $t$. We extend the Viterbi algorithm in the following manner to incorporate multiple feature vector information. The following equation estimate the most likely state sequence given an HMM $\lambda$ and a series of observations $O_t(l)$.

$$\begin{cases} \textit{Initilization}: \delta_1(i) = \pi_i \prod_{l=1}^{R} b_i(O_1(l)), \quad 1 \leqslant i \leqslant N, \\ \textit{Recursion}: \begin{cases} \delta_t(j) = \max_{1 \leqslant i \leqslant N} [\delta_{t-1}(i) a_{ij}] \prod_{l=1}^{R} b_j(O_t(l)), \\ \Psi_t(j) = \arg \max_{1 \leqslant i \leqslant N} [\delta_{t-1}(i) a_{ij}], \end{cases} \end{cases} \tag{30}$$

where $\pi_i$ is the initial probability of being in state $i$. The model parameters are adjusted in such a way that they can maximize the likelihood function for classifying actions using the given set of training data.

$$\lambda = \arg \max_{\lambda_a \in allActions} P(O|\lambda_a). \tag{31}$$

In this equation, $O$ represents the unknown feature vector sequence of an unknown action and $\lambda_a$ represents one MDHMM from the set of all known actions. The classifier recognizes the performed action by finding the model $\lambda$ with the highest conditional probability. Therefore, the values of $P(O|\lambda_a)$ for all models have to be computed. This is done by using a Viterbi decoder as mentioned earlier. Before this task, all model parameters have to be estimated first. HMM with discrete probability or continuous density can be used. Discrete symbolizing

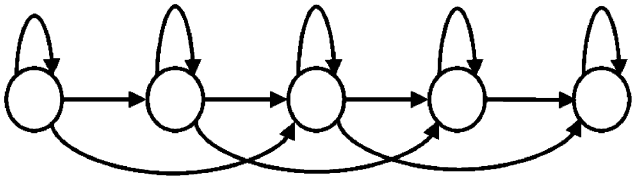

Fig. 8. Left–right HMM structure for an action.

via vector quantization induces degradation of performance. Therefore, continuous HMMs are used for action recognition. Continuous observation density is calculated by

$$b_j(o) = \sum_{k=1}^{M} c_{jk} \mathcal{N}(o, \mu_{jk}, \Sigma_{jk}), \tag{32}$$

where $o$ is the observation vector being modeled, $c_{jk}$ is the mixture coefficient for the $k$th mixture in state $j$ and $\mathcal{N}$ is the Gaussian. The multivariate Gaussian, $\mathcal{N}(o; \mu_{jk}, \Sigma_{jk})$ with mean vector $\mu$ and covariance matrix $\Sigma$, is the following:

$$\mathcal{N}(o; \mu, \Sigma) = \frac{1}{\sqrt{2\pi^{(n)}|\Sigma|}} e^{1/2(o-\mu)^{\mathrm{T}}(\Sigma^{-1})(o-\mu)}. \tag{33}$$

The maximum likelihood estimates of the average $\mu_j$ and covariance $\Sigma_j$ is $\hat{\mu}_j = \frac{1}{T}\sum_{t=1}^{\mathrm{T}} o_t$ and $\hat{\Sigma}_j = \frac{1}{T}\sum_{t=1}^{\mathrm{T}}(o_j - \mu_j)(o_j - \mu_j)^{\mathrm{T}}$, respectively. The left–right HMM with a strict left-to-right transition constraint and order structure is generally used to recognize action and speech, which is shown in Fig. 8, because of a temporal constraint in the action and speech patterns. The number of HMM states depends on the average action signal length, the complexity, and the variability of the pattern. We choose the number of key features of the action sequence as the number of states.

## 6. Experimental results and discussion

To illustrate the concepts, procedures, and human action recognition based on HMMs, we performed experiments on image sequences of different actions in different viewing directions.

### 6.1. Databases

#### 6.1.1. KUGDB

The aim of the KUGDB [27] is to form a data set for "state-of-art" action recognition and gesture recognition. The KUGDB contains 14 representative full body actions in the daily life of 20 performers. In the database, all the performers are elderly persons (both male and female) with ages ranging from 60 to 80. The database contains 3D motion data and three pairs of stereo video data taken at three different directions for each action using 3D motion capture devices and stereo cameras. The 2D data consist of both video data and 2D silhouette data. The data set includes three views, 0°, −45°, and +45°, respectively. The image sequences have 320 × 240 pixel resolutions and a frame rate of 30 frames per second. The testing set can be any arbitrary view. We use seven actions for training and testing purposes, which are shown in Fig. 9.

#### 6.1.2. KTHDB

The KTHDB [6] is one of the largest databases with sequences of human actions taken over different scenarios. The database contains six types of human actions, performed several times by 25 subjects in four scenarios given in Table 1. The sample images are shown in Fig. 10. The image sequences have 160 × 120 pixel resolutions and a frame rate of 25 frames per second. In KTHDB, there are $25 \times 6 \times 4 = 600$ video files for each combination of all subjects, actions and scenarios. Each file contains about four subsequences and used as a sequence in the experiments. Table 2 presents the sample frame number for each action and scenario. For more convenience, we show the minimum and maximum number of frames for any action.

### 6.2. Estimation of the temporal boundary

We estimate the period or duration by correlation or using the variation in pixel distribution in the silhouette image sequences. Let us consider that $p$ is the period. Therefore, the periodicity relationship becomes, $f(t+p) = f(t)$, where $f(t)$ is the motion of a point, or energy of an image at any time $t$. A non-periodic function is one that has no such period, instead we use the duration of action.

The brief algorithm for detecting period (or duration) is as follows: firstly, estimate the silhouette energy or correlation of image sequence. Secondly, apply smoothing operation to the similarity plot for periodic action and extract peak points. For non-periodic action, we apply non-maxima suppression method and make decision to extract the peak points (starting point and ending point). We choose multi-scale non-maxima window size for selecting the peak points, where non-maxima values are chosen arbitrarily. Now, the period is given by the difference between starting point and ending point as illustrated in Fig. 11. From the plots shown in Fig. 11, the approximate range of the period, the starting state, ending state, and phase variations are estimated.

### 6.3. Classification results

Fig. 12 shows the confusion matrix of action recognition using MDHMMs where we use shape flow, CLG motion flow, and combined features. Each column in the Figure represents the best match for each test sequence. We use 11 subjects, 7 actions, and 3 views variation for testing. At first, we recognized actions for shape flow and motion flow in arbitrary view directions. Then, we recognized actions for combined features in the arbitrary views. As can be seen, there is a clear separation among different kinds of actions. Among 7 actions to be classified, the most confusion occurs between walking and running. The confusion also occurs between 'sitting on the floor' and 'lying down on the floor' as well as 'sitting on a chair' and 'bowing'. This may occur due to the strong similarity between the action pair. The correct recognition rate (*CRR*) for

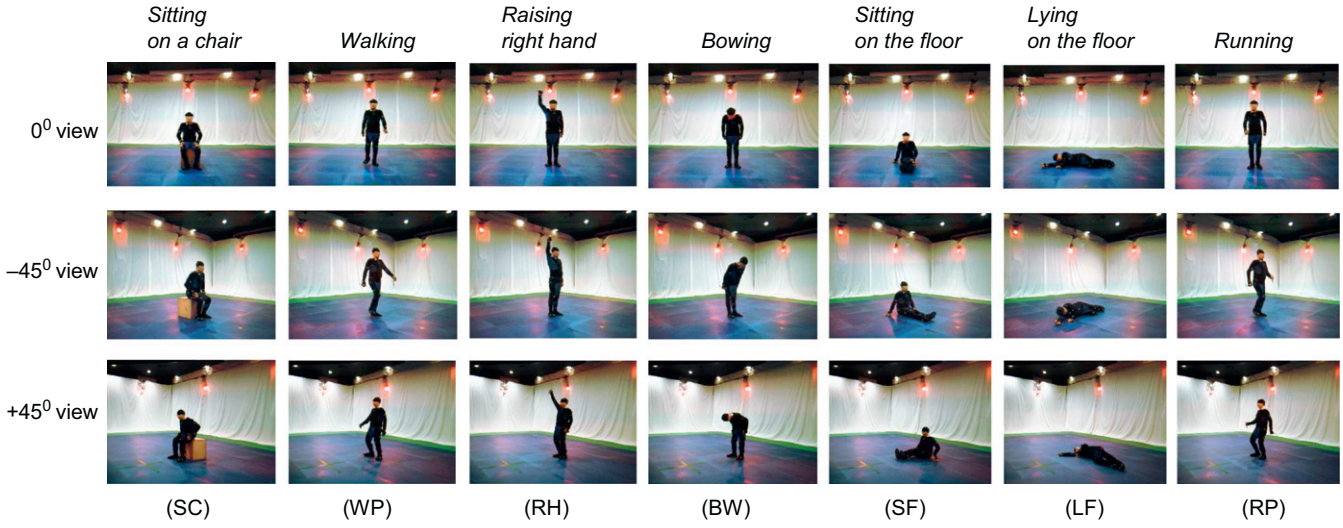


Fig. 9. Example images from KUGDB at three viewing directions. The bottom symbols represent the actions on the top.

Table 1
Scenarios in KTHDB

| Scenario | *s*1 | *s*2 | *s*3 | *s*4 |
|---|---|---|---|---|
| Description | Outdoors | Outdoors with scale variation | Outdoors with different cloths | Indoor |

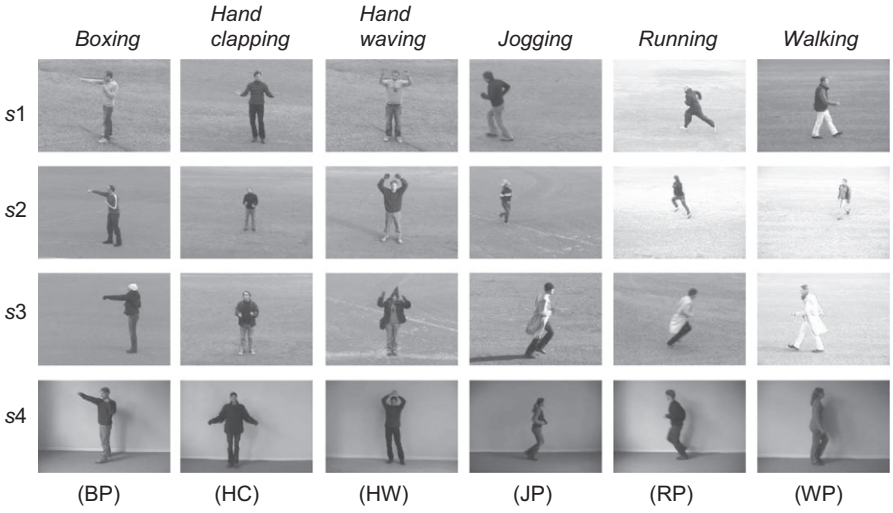


Fig. 10. Example images from KTHDB at four scenarios. The bottom symbols represent the actions on the top.

Table 2
Number of frames in each subsequence of each scenario

| Action | BP | | HC | | HW | | JP | | RP | | WP | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | min | max | min | max | min | max | min | max | min | max | min | max |
| *s*1 | 70 | 120 | 80 | 128 | 110 | 140 | 42 | 58 | 35 | 38 | 64 | 104 |
| *s*2 | 78 | 140 | 64 | 130 | 82 | 136 | 62 | 122 | 45 | 92 | 98 | 180 |
| *s*3 | 68 | 120 | 85 | 142 | 86 | 136 | 48 | 70 | 30 | 65 | 74 | 128 |
| *s*4 | 88 | 130 | 85 | 158 | 104 | 138 | 64 | 80 | 40 | 62 | 90 | 115 |

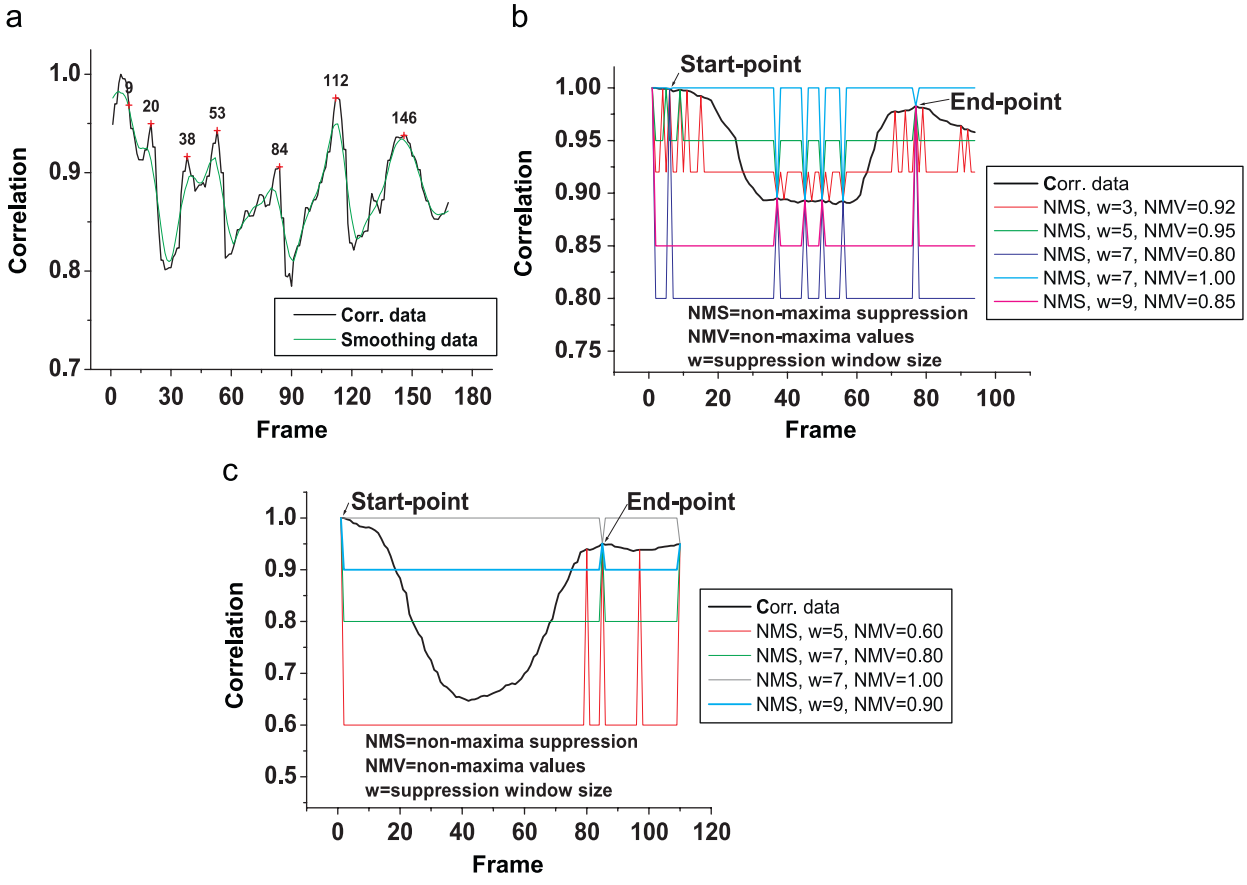


Fig. 11. Periodicity (or duration) detection from silhouette image sequences (KUGDB). (a) Running with multiple cycles $(t_s, t_e) = \{(20, 53), (53, 84), (84, 112), (112,146)\}$ with smoothing. (b) Raising the right-hand action ($t_s = 6$, $t_e = 77$). (c) Bowing action ($t_s = 1$, $t_e = 85$).

a

| | SC | WP | RH | BW | SF | LF | RF |
|---|---|---|---|---|---|---|---|
| SC | 0.82 | 0.06 | 0 | 0.06 | 0.06 | 0 | 0 |
| WP | 0 | 0.81 | 0.06 | 0 | 0 | 0 | 0.13 |
| RH | 0 | 0.13 | 0.78 | 0 | 0 | 0 | 0.09 |
| BW | 0.06 | 0 | 0 | 0.94 | 0 | 0 | 0 |
| SF | 0 | 0.06 | 0 | 0 | 0.81 | 0.13 | 0 |
| LF | 0.03 | 0 | 0 | 0 | 0.13 | 0.84 | 0 |
| RF | 0 | 0.16 | 0.06 | 0 | 0 | 0 | 0.78 |

b

| | SC | WP | RH | BW | SF | LF | RF |
|---|---|---|---|---|---|---|---|
| SC | 0.76 | 0.06 | 0 | 0.12 | 0.06 | 0 | 0 |
| WP | 0 | 0.75 | 0.06 | 0 | 0 | 0 | 0.19 |
| RH | 0 | 0.13 | 0.69 | 0 | 0 | 0 | 0.19 |
| BW | 0.06 | 0 | 0 | 0.88 | 0 | 0 | 0.06 |
| SF | 0 | 0.06 | 0 | 0 | 0.81 | 0.13 | 0 |
| LF | 0 | 0 | 0 | 0.06 | 0.22 | 0.72 | 0 |
| RF | 0.06 | 0.16 | 0.06 | 0 | 0 | 0 | 0.72 |

c

| | SC | WP | RH | BW | SF | LF | RF |
|---|---|---|---|---|---|---|---|
| SC | 0.85 | 0 | 0 | 0.09 | 0.06 | 0 | 0 |
| WP | 0 | 0.89 | 0 | 0 | 0 | 0 | 0.11 |
| RH | 0 | 0.06 | 0.81 | 0 | 0 | 0 | 0.13 |
| BW | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| SF | 0 | 0 | 0 | 0 | 0.88 | 0.13 | 0 |
| LF | 0 | 0 | 0 | 0 | 0.09 | 0.91 | 0 |
| RF | 0 | 0.13 | 0.03 | 0 | 0 | 0 | 0.84 |

Fig. 12. Confusion matrices for the KUGDB. (a) Shape flow feature. (b) CLG optic flow feature. (c) Combined shape and CLG flow features.

a

| | BP | HC | HW | JP | RP | WP |
|---|---|---|---|---|---|---|
| BP | 1 | 0 | 0 | 0 | 0 | 0 |
| HC | 0.03 | 0.97 | 0 | 0 | 0 | 0 |
| HW | 0 | 0.04 | 0.96 | 0 | 0 | 0 |
| JP | 0 | 0 | 0 | 0.88 | 0.04 | 0.08 |
| RP | 0.04 | 0 | 0 | 0.17 | 0.75 | 0.04 |
| WP | 0 | 0 | 0 | 0.1 | 0.05 | 0.85 |

b

| | BP | HC | HW | JP | RP | WP |
|---|---|---|---|---|---|---|
| BP | 0.96 | 0.04 | 0 | 0 | 0 | 0 |
| HC | 0.07 | 0.89 | 0.04 | 0 | 0 | 0 |
| HW | 0 | 0.04 | 0.92 | 0 | 0.04 | 0 |
| JP | 0 | 0 | 0 | 0.85 | 0.11 | 0.04 |
| RP | 0 | 0.05 | 0 | 0.14 | 0.68 | 0.14 |
| WP | 0 | 0 | 0.02 | 0.12 | 0.07 | 0.79 |

c

| | BP | HC | HW | JP | RP | WP |
|---|---|---|---|---|---|---|
| BP | 1 | 0 | 0 | 0 | 0 | 0 |
| HC | 0.04 | 0.96 | 0 | 0 | 0 | 0 |
| HW | 0 | 0.07 | 0.93 | 0 | 0 | 0 |
| JP | 0 | 0 | 0 | 0.88 | 0.08 | 0.04 |
| RP | 0 | 0.04 | 0 | 0.08 | 0.79 | 0.08 |
| WP | 0 | 0 | 0 | 0.1 | 0.07 | 0.83 |

Fig. 13. Confusion matrices for the KTHDB using combined shape and CLG motion flow features: (a) $s1$ scenario. (b) $s2$ scenario. (c) $s3$ scenario.

any action is calculated by

$$CRR(\%) = C/N \times 100, \tag{34}$$

where $C$ is the total number of correct recognition sequences while $N$ is the number of total action sequences. The CRRs are 82.57%, 76.14%, and 88.29% for shape flow, CLG motion flow,

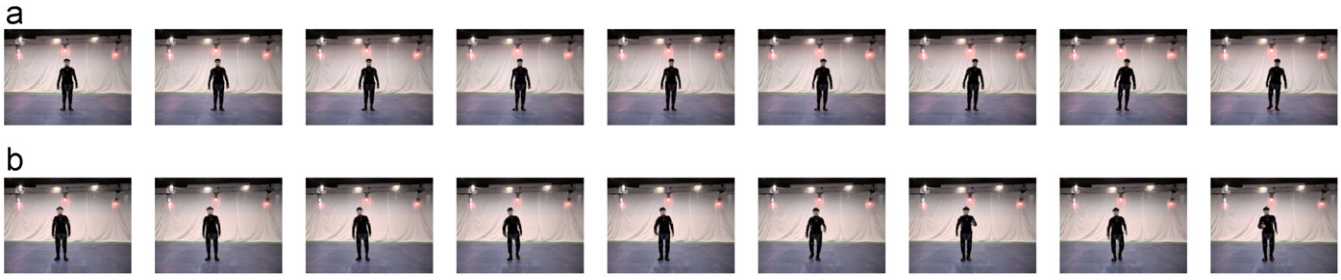


Fig. 14. Image sequences of "walking" and "running" from the front view (KUGDB).

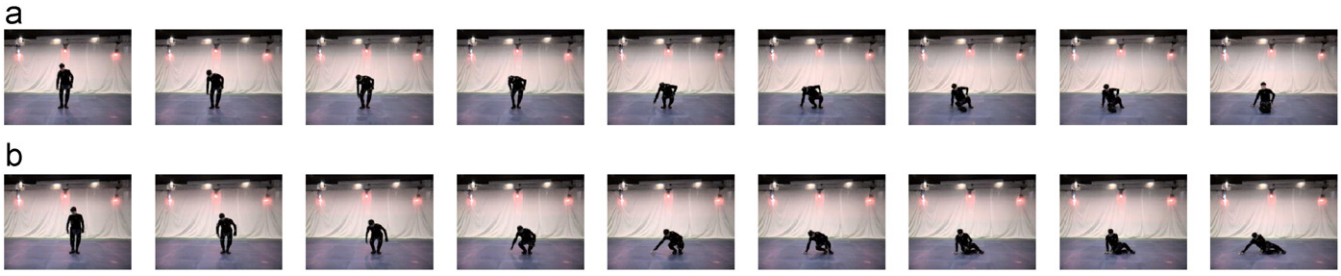


Fig. 15. Image sequences of "sitting on the floor" and "lying down on the floor" from the front view (KUGDB).

and combined flow in the arbitrary view directions. We also have tested our approach by using the KTHDB, since it is one of the largest human action databases and several researchers have used this database. We use 11 subjects, 6 actions, and 4 scenarios for testing and 8 subjects for training. The confusion matrices for six-class actions for the KTHDB are shown in Fig. 13 using combined shape and CLG optic flow features. We use each scenario and all scenarios for recognizing actions for the same data mentioned in Fig. 13. The most confusion occurs between jogging and running as well as jogging and walking, although it varies at different scenarios. The recognition accuracies are 90.17%, 84.83%, 89.83%, and 85.67% for scenarios $s1$, $s2$, $s3$, and $s4$, respectively. We use the features which are invariant to translation, rotation, and scale of the person, therefore, scenario $s2$, which includes the scale-varying action, has very little effect on recognition performance.

In the testing phase of the experiment, we find that some sequences are misclassified, such as walking and running, sitting on the floor and lying on the floor. These sequences are checked manually, and it is found that these image sequences are taken from the front (0°) view. These situations are shown in Fig. 14. They may be a result of the high degree of similarity between walking and running in the image in the front view. Moreover, all performers are elderly people in the KUGDB and naturally their walking and running motions are similar. In the case of "sitting on the floor" and "lying down on the floor", it is found that several image frames have strong similarity at the middle stage of the action, which is shown in Fig. 15. In the front view, the strong similarity occurs between walking and running, so recognition rate is lower. But in the side view, it is easier to distinguish actions.

We have shown the timing data for our method for extracting both shape and motion flows. In case of shape features, we used the invariant Zernike moment which is robust to noise but the computation cost of Zernike moment is expensive due to its orthogonal property. Our measurement is implemented in C/C++ on a 1.70 GHz Pentium IV PC. Tables 3 and 4 present the timing data of feature extraction procedure.

Table 3
Range of timing data for feature extraction (KUGDB, image size = 320 × 240)

| Feature | Shape | Motion | Combined |
|---|---|---|---|
| Without Zernike moment (s/frame) | 0.12–0.14 | 0.26–0.32 | 0.40–0.50 |
| With Zernike moment (s/frame) | 2.65–2.92 | 0.26–0.32 | 2.80–2.84 |

Table 4
Range of timing data for feature extraction (KTHDB, image size = 160 × 120)

| Feature | Shape | Motion | Combined |
|---|---|---|---|
| Without Zernike moment (s/frame) | 0.06–0.07 | 0.06–0.07 | 0.09–0.11 |
| With Zernike moment (s/frame) | 0.62–0.67 | 0.06–0.07 | 0.69–0.82 |

### 6.4. Comparison

At first, we compare our human action recognition approach with other previous researches, along with features selection, view direction, and recognition rate, irrespective of the captured video sequences, as given in Table 5. It is difficult to compare these approaches, since the data sets and environments are different. However, the results can give a general overview and comparison of some approaches in action recognition. The important declaration of this work is, "we recognize human action from any arbitrary view rather than any specific view with

Table 5
Comparison results of action recognition with some previous researches

| Researches | Action | Feature type | View | Recognition rate |
|---|---|---|---|---|
| Ali et al. [17] | 7 | Angle of three body components | Profile view | 78.8 |
| Sun et al. [10] | – | Affine and optic flows | Single view | 90.0 |
| Masaud et al. [12] | 8 | Motion | Front-parallel | 92.8 |
| Yacoob et al. [13] | 4 | Parametric motion | Diagonal | 82.0 |
| Zobl et al. [14] | 6 | Global motion | View-dependent | 66.0 |
| Sheikh et al. [5] | 4 | Subspace angle | View-independent | – |
| Schüldt et al. [6] | 6 | Local space-time | Multiple scenarios | 71.72 |
| Our approach | 7 | Motion and shape flows (KUGDB) | View-independent | 88.29 |
| Our approach | 6 | Motion and shape flows (KTHDB) | Multiple scenarios | 88.33 |

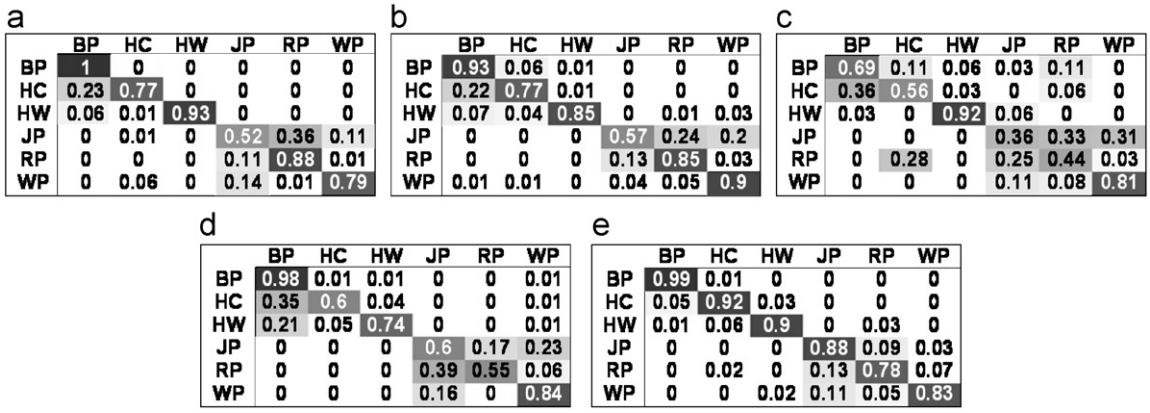


Fig. 16. Comparison of confusion matrices for the KTHDB. (a) Niebles's method; (b) Dollár's method; (c) Ke's method; (d) Schüldt's method; (e) our method.

Table 6
Comparison results of the action recognition using KTHDB

| Method | Recognition accuracy | Scenarios |
|---|---|---|
| Niebles et al. [9] | 81.50 | $s1$+$s2$+$s3$+$s4$ |
| Dollár et al. [7] | 81.17 | $s1$+$s2$+$s3$+$s4$ |
| Schüldt et al. [6] | 71.72 | $s1$+$s2$+$s3$+$s4$ |
| Ke et al. [8] | 62.96 | $s1$+$s2$+$s3$+$s4$ |
| Our method | 88.33 | $s1$+$s2$+$s3$+$s4$ |
| Schüldt et al. [6] | 62.33 | $s2$ |
| Our method | 84.83 | $s2$ |

additional variability selection". Secondly, we compare our works with some state-of-art action recognition approaches by using the same database.

In addition, we compare our method against Schüldt et al.'s [6] in classifying periodic actions. We use the same training and test sequences as in their paper, which contains eight people in the training set and nine in the testing. Each person repeats six actions in each of four scenarios. In addition with [6], we compare our results with the best results from [7,9,8] using KTHDB. Our results by combined motion and shape flow are on par with their results obtained by local space time features. The comparison of the confusion matrices ($s1$+$s2$+$s3$+$s4$ scenarios) is shown in Fig. 16. Most confusion occurs between jogging and running, jogging and walking, and hand clapping and boxing. In our approach, confusion occurs between jogging and running as well as jogging and walking. The overall comparison of different methods is listed in Table 6. Compared to the mentioned researches, our approach yields better recognition results.

## 7. Conclusions and future research

This paper addressed robust human action recognition from multiple view image sequences by using combined shape flow and CLG motion flow. We also considered some sources of variability that affect action recognition. This variability includes the view-directional variation, anthropometry variation, scale variation, and phase change of action. Based on the combined features, a set of MDHMMs were built for the mentioned actions, to represent each action from multiple views or each scenario and enable recognizing from arbitrary views. We first showed that combined feature information increased the recognition than individual feature by using KUGDB and then we recognized action by using only the combined features. We recognized different daily human actions successfully in the indoor environment as well as in the outdoor environment. The action recognition rate compared with some previous researches is not much higher but it was shown that we recognized actions from multiple views rather than a set view. Moreover, we recognized action at a higher rate with the same kind of data. The recognition rate of combined features is higher than the rate obtained by either shape or motion feature using the KUGDB.

This result showed that our algorithm is robust to variations in view and duration. Our proposed method for action recognition is flexible since it can be adapted to practical applications of human movement, human action recognition, and so on. This is different from other shape-based or motion-based variation approaches where analysis is done at a single level. We included the features of silhouettes and original images when a person performs action in different speed variation, change of phase variation, i.e. the starting and ending phase variation of actions. This enforces the robustness of the action recognition.

Basically, action recognition applications require the development of systems which are fast, can handle a variety of actions, need a limited number of parameters, need as fast as possible learning stage, and are robust to environment variation. In our approach, we tried to full-fill the requirements of such a system. Despite of robustness of the system, we faced the problem of recognition in a few actions which occurred in the fronto-normal view. The current complexity of our experiment is the optimal number of features, and styles of action, although the selected features are robust for action recognition. Our future work includes the interaction of multi-view learning using the adaptable hidden Markov model with complex multiple human actions.

## Acknowledgment

This research was supported by the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Ministry of Commerce, Industry and Energy of Korea.

## References

[1] T.B. Moeslund, A. Hilton, V. Krüger, A survey of advances in vision-based human motion capture and analysis, Comput. Vision Image Understanding 104 (2006) 90–126.

[2] W. Hu, T. Tan, L. Wang, S. Maybank, A survey on visual surveillance of object motion and behaviors, IEEE Trans. Syst. Man Cybern.—Part C: Appl. Rev. 34 (3) (2004) 334–352.

[3] D.M. Gavrilla, The visual analysis of human movement: a survey, Comput. Vision Image Understanding 73 (1) (1999) 82–98.

[4] J.K. Aggarwal, Q. Cai, Human motion analysis: a review, Comput. Vision Image Understanding 73 (3) (1999) 428–440.

[5] Y. Sheikh, M. Shah, M. Shah, Exploring the space of a human action, in: Proceedings of the IEEE International Conference on Computer Vision, October 2005, pp. 144–149.

[6] C. Schüldt, I. Laptev, B. Caputo, Recognizing human actions: a local SVM approach, in: Proceedings of the IEEE International Conference ICPR, vol. 3, 2004, pp. 32–36. The KTHDB, ⟨http://www.nada.kth.se/cvap/actions/⟩.

[7] P. Dollár, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal filters, in: Proceedings of the IEEE International Workshop VS-PETS, 2005, pp. 65–72.

[8] Y. Ke, R. Sukthankar, M. Hebert, Efficient visual event detection using volumetric features, in: Proceedings of IEEE International Conference on ICCV, 2005, pp. 166–173.

[9] J.C. Niebles, H. Wang, L. Fei-Fei, Unsupervised learning of human action categories using spatial–temporal words, Proc. BMVC 3 (2006) 1249–1258.

[10] X. Sun, C. Chen, B.S. Manjunath, Probabilistic motion parameter models for activity recognition, in: Proceedings of IEEE International Conference on Pattern Recognition, vol. 1, Quebec City, Canada, August 2002, pp. 443–446.

[11] J. Ben-Arie, Z. Wang, P. Pandit, S. Rajaram, Human activity recognition using multidimensional indexing, IEEE Trans. Pattern Anal. Mach. Intell. 24 (8) (2002) 1091–1104.

[12] O. Masoud, N. Papanikolopoulos, Recognizing human activities, in: Proceedings of IEEE Conference on Advanced Video and Signal Based Surveillance, Florida, USA, July 2003, pp. 157–162.

[13] Y. Yacoob, M.J. Black, Parameterized modeling and recognition of activities, Comput. Vision Image Understanding 73 (2) (1999) 232–247.

[14] M. Zobl, F. Wallhoff, G. Rigoll, Action recognition in meeting scenarios using global motion features, in: Proceedings of IEEE International Workshop on PETS, 2003.

[15] S. Seitz, C. Dyer, View invariant analysis of cyclic motion, Int. J. Comput. Vision 25 (1997) 231–251.

[16] M. Ahmad, S.-W. Lee, HMM-based human action recognition using multiview image sequences, in: Proceedings of the IEEE International Conference on Pattern Recognition, Hong Kong, vol. 1, August 2006, pp. 263–266.

[17] A. Ali, J.K. Aggarwal, Segmentation and recognition of continuous human activity, in: Proceedings of IEEE Workshop on Detection and Recognition of Events in Video, Canada, July 2001, pp. 28–35.

[18] S. Niyogi, E. Adelson, Analyzing and recognizing walking figures in XYT, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 1994, pp. 469–473.

[19] R. Collins, R. Gross, J. Shi, Silhouette based human identification using body shape and gait, in: Proceedings of the International Conference on Automatic Face and Gesture Recognition, 2002, pp. 351–356.

[20] J. Foster, M. Nixon, A. Prugel-Bennett, Automatic gait recognition using area based matrices, Pattern Recognition Lett. 24 (2003) 2489–2497.

[21] I. Cohen, H. Li, Inference of human postures by classification of 3D human body shape, in: Proceedings of IEEE International Workshop on Analysis and Modeling of Faces and Gestures, 2003, pp. 74–81.

[22] S. Carlsson, J. Sullivan, Action recognition by shape matching to key frames, Proceedings of IEEE CS Workshop on Models versus Exemplars in Computer Vision, Florida, USA, 2002, pp. 263–270.

[23] A. Veeraraghavan, A.K. Roy-Chowdhury, R. Chellappa, Matching shape sequences in video with applications in human movement analysis, IEEE Trans. Pattern Anal. Mach. Intell. 27 (12) (2005) 1896–1909.

[24] A.F. Bobick, J.W. Davis, The recognition of human movement using temporal templates, IEEE Trans. Pattern Anal. Mach. Intell. 23 (3) (2001) 257–267.

[25] C. Rao, M. Shah, View-invariance in action recognition, in: Proceedings of IEEE CS Conference on Computer Vision and Pattern Recognition, Hawaii, USA, December 2001, pp. 316–323.

[26] V. Parameswaran, R. Chellappa, View invariants for human action recognition, in: Proceedings of IEEE CS Conference on Computer Vision and Pattern Recognition, vol. 2, 2003, pp. 613–619.

[27] B.-W. Hwang, S. Kim, S.-W. Lee, A full-body gesture database for human gesture analysis, Int. J. Pattern Recognition Artif. Intell. 21 (6) (2007) 1069–1084.

[28] T. Hoprasert, D. Harwood, L.S. Davis, A statistical approach for real-time robust background subtraction and shadow detection, in: Proceedings of the 7th IEEE International Conference on Computer Vision, Frame Rate Workshop, Greece, September 1999, pp. 1–19.

[29] M.-K. Hu, Visual pattern recognition by moment invariants, IRE Trans. Inf. Theory IT-8 (1962) 179–187.

[30] A. Khotanzad, Y.H. Hong, Invariant image recognition by Zernike moments, IEEE Trans. Pattern Anal. Mach. Intell. 12 (5) (1990) 489–497.

[31] C.-W. Chong, P. Raveendran, R. Mukundan, Translation invariants of Zernike moments, Pattern Recognition 36 (2003) 1765–1773.

[32] C.-W. Chong, P. Raveendran, R. Mukundan, The scale invariants of pseudo-Zernike moments, Pattern Anal. Appl. 6 (2003) 176–184.

[33] A. Bruhn, J. Weickert, C. Schnörr, Lucas/Kanade meets Horn/Schunck: combining local and global optic flow methods, Int. J. Comput. Vision 61 (3) (2005) 211–231.

[34] B.K.P. Horn, B.G. Schunck, Determining optical flow, Artif. Intell. 17 (1981) 185–203.

[35] B. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, in: Proceedings of International Joint Conference on Artificial Intelligence, 1981, pp. 674–679.
[36] C. Bregler, Learning and recognizing human dynamics in video sequences, in: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 1997, pp. 568–574.
[37] J.A. Montero, L.E. Sucar, Feature selection for visual gesture recognition using hidden Markov models, in: Proceedings of the Fifth International Conference in Computer Science, 2004.
[38] J. Yamato, J. Ohya, K. Ishii, Recognizing human action in time-sequential images using hidden Markov model, in: Proceedings of IEEE International Conference on CVPR, 1992, pp. 379–385.
[39] M. Brand, M. Oliver, A. Pentland, Coupled hidden Markov model for complex action recognition, in: Proceedings of IEEE CS Conference on CVPR, 1997, pp. 994–999.
[40] R.V. Babu, B. Anantharaman, K.R. Ramakrishnan, S.H. Srinivasan, Compressed domain action classification using HMM, Pattern Recognition Lett. 23 (2002) 1203–1213.
[41] H. Bunke, T. Caelli, HMMs Applications in Computer Vision, World Scientific, Singapore, 2001.
[42] B.U. Toreyin, Y. Dedeoglu, A.E. Cetin, HMM based falling person detection using both audio and video, Lecture Notes in Computer Science, vol. 3766, Springer, Berlin, 2005, pp. 211–220.
[43] N.P. Cuntoor, B. Yegnanarayana, R. Chellappa, Interpretation of state sequences in HMM for activity representation, in: Proceedings of IEEE ICASSP'05, 2005, pp. 709–712.
[44] R. Lawrence, A. Rabiner, Tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE 77 (2) (1989) 257–286.
[45] J. Yang, Y. Xu, C.S. Chen, Hidden Markov model approach to skill learning and its application in telerobotics, IEEE Trans. Robotics Autom. 10 (5) (1994) 621–631.

**About the author**—MOHIUDDIN AHMAD received his B.S. degree with Honors in Electrical and Electronic Engineering from Chittagong University of Engineering and Technology, Bangladesh and his M.S. degree in Electronics and Information Science from Kyoto Institute of Technology of Japan in 1994 and 2001, respectively. He is currently a Ph.D. candidate in the Department of Computer Science and Engineering, Korea University, Korea. From August 1995 to October 1998, he served as a lecturer at Khulna University of Engineering and Technology, Bangladesh. In June 2001, he joined the same Department as an assistant professor. His research interests include human action recognition, modeling, image processing and their applications in computer vision and the pattern recognition related fields.

**About the author**—SEONG-WHAN LEE received his B.S. degree in Computer Science and Statistics from Seoul National University, Seoul, Korea, in 1984, and his M.S. and Ph.D. degrees in computer science from KAIST in 1986 and 1989, respectively. From February 1989 to February 1995, he was an assistant professor in the Department of Computer Science at Chungbuk National University, Cheongju, Korea. In March 1995, he joined the faculty of the Department of Computer Science and Engineering at Korea University, Seoul, Korea, as an associate professor, and he is now a full professor. He was the winner of the Annual Best Paper Award of the Korea Information Science Society in 1986. He obtained the First Outstanding Young Researcher Award at the Second International Conference on Document Analysis and Recognition in 1993, and the First Distinguished Research Professor Award from Chungbuk National University in 1994. He also obtained the Outstanding Research Award from the Korea Information Science Society in 1996. He also received an Honorable Mention of the Annual Pattern Recognition Society Award for an outstanding contribution to the Pattern Recognition Journal in 1998. He is a fellow of International Association for Pattern Recognition, a senior member of the IEEE Computer Society and a life member of the Korea Information Science Society. He has published more than 200 publications in these areas in international journals and conference proceedings, and has authored 10 books. His research interests include pattern recognition, computer vision, and biometrics.